J. Clin. Chem. Clin. Biochem.
Vol. 22, 1984, pp. 431–445

Comparison of Several Regression Procedures for Method Comparison Studies and Determination of Sample Sizes

Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part II

By *H. Passing*

Abt. für Praktische Mathematik, Hoechst AG, Frankfurt am Main 80, and

W. Bablok

Allg. Biometrie, Boehringer Mannheim GmbH, Mannheim

(Received October 18, 1983/February 27, 1984)

Summary: In part I of this series (*H. Passing & W. Bablok* (1983), J. Clin. Chem. Clin. Biochem. *21*, 709–720) we described a new biometrical procedure for the evaluation of method comparison studies. In part II we now discuss its properties and compare them with those of other established procedures by means of a simulation study. We demonstrate that the reliability of the results not only depends on the sample size but also on the sampling distribution, the precision of the methods, and the concentration range covered by the samples. Linear regression and principal component procedures are either inadequate or not as reliable as our new procedure. The appropriate sample size is discussed and recommendations are given.

Vergleich verschiedener Regressionsverfahren für Methodenvergleichsstudien und Bestimmung von Stichprobenumfängen

Anwendung von linearen Regressionsverfahren bei Methodenvergleichsstudien in der Klinischen Chemie, Teil II

Zusammenfassung: Im ersten Teil unserer Arbeit (*H. Passing & W. Bablok* (1983), J. Clin. Chem. Clin. Biochem. *21*, 709–720) haben wir ein neues biometrisches Verfahren zur Auswertung von Methodenvergleichen vorgestellt. Im zweiten Teil nun werden seine Eigenschaften untersucht und im Rahmen einer Simulationsstudie mit denen anderer, bereits etablierter Verfahren verglichen. Wir zeigen, daß die Zuverlässigkeit der Auswertungsergebnisse nicht nur vom Stichprobenumfang, sondern auch von der Stichprobenverteilung, der Präzision der Methoden und dem Konzentrationsbereich der Proben abhängt. Lineare Regression und die Hauptkomponenten-Verfahren sind entweder unzulänglich oder nicht so zuverlässig wie das neue Verfahren. Die Abhängigkeit des Stichprobenumfanges von den Randbedingungen wird diskutiert, und Empfehlungen werden gegeben.

Contents

1. Introduction

In part I of our paper (1) we described a new statistical linear regression procedure which can be employed in the evaluation of method comparison studies. We showed its theoretical advantages compared with other statistical procedures and presented an example with real data to support our arguments.

In this paper we present the comparison of our procedure with other established procedures in order to demonstrate its merits. Further we shall expose the strong dependence of the appropriate sample size on certain properties which are inherent in the analytical methods for which a comparison should be performed.

The experimental design consists of drawing n independent samples from a population and measuring the analyte in question with each of the two methods. The evaluation usually consists of fitting a straight line $Y = \alpha + \beta X$ to the data and in testing the hypotheses $\beta = 1$ and $\alpha = 0$. The evaluation procedures discussed here differ with regard to the estimators, the meaning of β and α, and the tests of the hypotheses. The statistical models on which an evaluation can be based are discussed in part I. The notation and meaning of X and Y will be the same as in part I.

In this context we should like to point out that the problem of determining the parameters of a linear equation by which one method is transformed into the other is a different one. It deserves a separate treatment and will be the topic of part III of our paper.

2. Evaluation Procedures for Method Comparison

The evaluation procedures discussed in this paper are divided in two groups, one which is invariant with regard to the assignment of the methods to X and Y, and one which is not. A procedure is invariant, if after interchanging X and Y the respective estimators of β and α can be transformed into each other, and if the test of the hypotheses $\beta = 1$ and $\alpha = 0$ gives the same results under both assignments.

2.1 Invariant procedures

The invariant procedures asume that the variation of the observed values x_i and y_i has two independent sources: one is the variation within the population of all possible samples, the other one is the measurement error within each sample. This leads to the partition

$$x_i = x_i^* + \xi_i \quad \text{and} \quad y_i = y_i^* + \eta_i,$$

where x_i^* and y_i^* are the expected values within the i-th sample and ζ_i and η_i are the measurement errors. The procedures assume the structural relationship

$$y_i^* = \alpha + \beta x_i^*$$

between the expected values.

Procedure P_1: This is our new procedure as described in part I. The estimators of β and α are both medians. The usual statistics, such as mean, standard deviation and correlation coefficient are not required. The hypothesis tests do not ask for any specific distributional assumptions.

Procedure P_2: Standardized principal component analysis (2). Its estimators are based on means and standard deviations; the tests require certain distributional properties of the data.

Procedure P_3: Principal component analysis (2, 3). The theoretical background is identical to that of P_2 and the test of the hypothesis $\beta = 1$ is the same; the estimators, however, involve in addition the coefficient of correlation.

2.2 Procedures which are not invariant

These procedures assume that one variable is free of random variation implying that it is fixed.

The underlying statistical models are given by

$$y_i = \alpha + \beta x_i + \eta_i \quad \text{with fixed } x_i, \text{ or}$$
$$x_i = A + B y_i + \xi_i \quad \text{with fixed } y_i.$$

2.2.1 *Procedures assuming X to be fixed*

Procedure P_4: *Theil*'s procedure (4): This is similar to P_1, in particular the estimator of β is also a medi-

an. The test of the hypothesis $\beta = 1$ is distribution-free. *Theil* does not give an estimator for α; α may be estimated as in P_1.

Procedure P_5: This is the classical linear regression based on least squares (5). Its estimators are based on means, standard deviations and the correlation coefficient (as for P_2 and P_3); the tests require certain distributional properties of the underlying data.

2.2.2 *Procedures asuming Y to be fixed*

Procedure P_6 is identical to procedure P_4 and *procedure P_7* identical to procedure P_5, only the assignment of the methods to X and Y is interchanged. (If the arithmetical evaluation of a method comparison is carried out by a procedure which is not invariant then usually both P_4 and P_6, or P_5 and P_7 are calculated.)

3. The Simulation Model

If the same data set is evaluated by the procedures P_1 to P_7 the estimators of β and α usually yield different results. Moreover, the results of testing the hypotheses $\beta = 1$ and $\alpha = 0$ are not necessarily identical. Therefore it is desirable to know which procedure really gives the correct result.

Each of the 7 procedures is based on certain mathematical assumptions which are different or even contrary to each other. If the properties of the real data meet the assumptions of a particular procedure it will give a reliable result; otherwise the result may be biased. Therefore, a given data set may satisfy the assumptions of one procedure but not of the other one so that systematic differences between the results can be expected. If we restrict ourselves to the slope β – and β is the most important parameter in such an evaluation – then deviations may occur in two respects:

Firstly, an estimator b of β may be biased in so far that it achieves values which are systematically larger (or smaller) than β. Hence b would not estimate β but anything else resulting possibly in an erroneous judgement of the methods. Consequently an unbiased estimation of b in realistic situations is a desirable property.

Secondly, testing the hypothesis $\beta = 1$ for the estimated slope b may result in a significant difference even though $\beta = 1$ is true. If the mathematical assumptions of the test are met the probability of such a false positive result is restricted to the level γ (e.g. $\gamma = 5\%$). Otherwise the actual level – that is the true probability of obtaining a false positive result under the given circumstances – may be much higher than the nominal level of γ on which the test is performed. Consequently, an evaluation procedure becomes inappropriate if significant differences between the two methods would be found too frequently. Therefore, a second desirable property of a procedure is to achieve the level γ of probability in realistic situations.

It follows that a procedure gives the correct result if its actual level is about γ and if its estimator of the slope is unbiased.

Obviously both properties cannot be studied by evaluating real data sets, since the true relation between both methods is not known. They can, however, be judged by the results of simulation experiments (6). Here data sets describing a well defined "situation" are repeatedly generated to study the behaviour of P_1 to P_7 in detail. Our simulation is based on the structural relationship model which gives a reasonable description of reality. Since procedures P_4 to P_7 are frequently used in method comparison studies we have included these procedures in the simulation study. From the structural relationship model, it follows that the CV's of the methods can be defined from the variances σ_ξ^2 and σ_η^2 of their respective measurement errors. For ease of notation, however, we shall use CV_x and CV_y when we refer to the coefficient of variation of method X or method Y respectively.

Before we describe the details of the simulation model we state the following general assumptions:

- There is a linear relationship between method X and method Y.
- The measurements of X and Y are realisations of independent continuous bivariate variables.

Let $[c_u, c_0]$ be the range of measurements for the method assigned to X, and $[\beta c_u, \beta c_0]$ the corresponding range for the method assigned to Y. Let $c = \frac{c_0}{c_u}$ be the *common size* of both ranges with $1 < c < \infty$. A *large size* corresponding to $c \geq 8$ will be represented by $c = \infty$, a *medium size* corresponding to $4 \leq c < 8$ by $c = 4$. Further, a *small size* of $2 \leq c < 4$ will be modelled by $c = 2$ and a *very small size* of $1.25 \leq c < 2$ by $c = 1.25$. Since every method has a lower detection limit there is always $c_u > 0$.

As a measure of precision which is known before the start of a method comparison experiment, we use the coefficient of variation. In the simulation model we assume CV_x and CV_y to be constant over their concentration range[1]). This is much more realistic than

[1]) The CV is chosen here as a familiar measure of precision. In l.c. (7) it is shown that other measures may be more appropriate if the distribution of the measurement error is skew or has a kurtosis, but this property is without relevance in this context.

the usual assumption of constant standard deviations (2, 5). For completeness sake, we also studied the influence of non-constant CV's on the evaluation procedures. The magnitude of the CV's is not independent of the size c of the measurement range since methods for constituents with a very small biological range (as for instance electrolytes) require small CV's for the differentiation of measurements. Therefore, the CV's are varied independently of each other from 2% to 13% for a medium or large size c and from 2% to 10% for a small size c. In the case of a very small size the CV's are varied from 1% to 2%.

It can be shown that it is sufficient to take the interval $[\frac{1}{c}, 1]$ for X and the interval $[\frac{\beta}{c}, \beta]$ for Y as common range, whereby both CV's remain unchanged. This does not cause any loss of generality, but achieves independence from the real size of the measurement values.

In the simulation model, n samples are drawn from the interval $[\frac{1}{c}, 1]$; the i-th sample has expected values x_i^* and $y_i^* = \beta x_i^*$. (Since we study the estimation of β we assume a constant α and set it equal to zero).

The samples are generated from

- a uniform sampling distribution over $[\frac{1}{c}, 1]$, represented by equidistant x_i^*, or
- a skew sampling distribution over $[\frac{1}{c}, 1]$: to achieve this the interval is divided into 5 sections of equal length with equidistant x_i^* covering 5%, 50%, 30%, 10% and 5% of n. In this way the measurements are concentrated more in the left part of the range. This distribution corresponds to many real situations where usually samples come both from healthy and diseased persons.

The choice of predefined x_i^* from the uniform and skew sampling distribution actually leads to a functional relationship model. However, our first investigations demonstrated that the results from the simulation on the basis of a true structural relationship model (with random generation of x_i^*) did not differ from those of the functional relationship model. Considering the amount of computing time needed for the simulation we decided to use the less demanding functional relationship model. Besides, this model can be interpreted as a special case of the structural relationship model.

In order to estimate β reliably it would be optimal to have the samples located at the boundaries of the range as long as linearity is guaranteed. Obviously, this sampling distribution will be insufficient for practical reasons. However, the uniform distribution lies between this extreme and the usually skew sampling distribution and is attainable.

The expected values are distorted by independent measurement errors ξ_i and η_i giving "measurement values" $x_i = x_i^* + \xi_i$ and $y_i = y_i^* + \eta_i$. These errors correspond to the precision of the methods. Three types of distribution of measurement errors are considered:

- ξ_i and η_i are normally distributed.
- ξ_i and η_i both have a mixture of two normal distributions differing slightly from each other so that the resulting distributions of ξ_i and η_i look like a normal distribution; in particular they are symmetric.
- ξ_i and η_i both have a skew distribution with a positive kurtosis so that they differ essentially from a normal distribution.

The latter two distributions are chosen since it is well known (8, 9) that in general the measurement errors are not normally distributed. Therefore it is advisable to investigate several distributions.

So far the simulation does not allow for large differences between x_i and y_i that occur frequently in real experiments. We call such a pair (x_i, y_i) an extreme value. It may be caused by a difference in specificity or by susceptibility to interferences of the methods and should not be removed from the evaluation without any experimental reason. Hence it is necessary to consider extreme values in the simulation model; they are introduced to the data by changing some values of Y up to $\pm$ 50%.

In summing up we define one simulation step by the following parameters: c = common size of ranges, n = sample size, CV_x, CV_y, β, sampling distribution, distribution of measurement errors and number and position of extreme values.

In our basic simulation model n pairs (x_i, y_i) are generated for each choice of simulation parameters. The estimators $b_1, \ldots, b_7$ of the slope are calculated according to $P_1, \ldots, P_7$, and the hypothesis $\beta = 1$ is tested with respect to $P_1, \ldots, P_7$. These steps are performed 500 times for each choice of the simulation parameters. From the 500 b_i's the median med (b_i) is calculated. The bias of b_i is estimated (med $(b_i) - \beta$). The proportion of significant test results given by P_i is an estimator of the actual level of $= P_i$ if $\beta = 1$. The 7 procedures are compared with regard to their actual level and to their bias, if $\beta \neq 1$. The case $\beta \neq 1$ deserves a separate investigation. As the maximum

sample size we choose n = 90 since otherwise the simulation expenditure would be too large. The influence of extreme values is not investigated and only normally distributed measurement errors are considered. The probability of a significant test result if $\beta \neq 1$ is true, is called the *power*. It is tabulated for several parameter combinations. The power should be large if there is a relevant difference between β and 1. Therefore the user must define a relevant value, β_{rel}, which is adequate to his specific problem. Then, if β is larger than β_{rel} or less than $1/\beta_{rel}$ the power should be sufficiently large, say $\geq 80\%$. It is shown that this desirable property depends on the suitable choice of the sample size n. A list of such sample sizes is given.

Since it is difficult to assess the properties of an experimental data set with respect to the model assumptions of a regression procedure, we find it more advantageous to demonstrate how a procedure behaves if certain assumptions are not met.

4. Comparison of Procedures P_1 to P_7 if $\beta = 1$

4.1 The probabillity of a false positive test result if no extreme values are present

We demonstrate the results of the simulation study for the sample size n = 40. The probabilities obtained for a false positive test result are accurate up to $\pm 2\%$. A summary is given in table 1, where $CV_x \leq CV_y$ is assumed without loss of generality. The

Tab. 1. Actual probability of a false positive test result for n = 40 (γ = 5%); no extreme values are present.

P_1 = our new procedure
P_2 = standardized principal component analysis
P_3 = principal component analysis
P_4 = *Theil*'s procedure, X assumed to be fixed
P_5 = least squares linear regression, X assumed to be fixed
P_6 = *Theil*'s procedure, Y assumed to be fixed
P_7 = least squares linear regression, Y assumed to be fixed

			Samples: uniform Errors: normal						Samples: skew Errors: normal						Samples: uniform Errors: mixture of normals						Samples: uniform Errors: skew with kurtosis					
			Invariant		Not invariant				Invariant		Not invariant				Invariant		Not invariant				Invariant		Not invariant			
c	CV_x (%)	CV_y (%)	P_1	P_2 P_3	P_4	P_5	P_6	P_7	P_1	P_2 P_3	P_4	P_5	P_6	P_7	P_1	P_2 P_3	P_4	P_5	P_6	P_7	P_1	P_2 P_3	P_4	P_5	P_6	P_7
∞	2	2	5	6	4	7	5	7	5	17	5	17	6	18	4	5	4	5	4	5	5	6	5	7	6	7
	2	5	6	8	6	8	6	11	6	17	6	18	9	20	6	6	7	7	7	8	6	16	6	10	7	23
	2	7	7	10	6	8	10	16	6	19	6	15	10	29	5	8	6	7	6	13	7	17	6	9	8	30
	5	7	6	8	7	9	9	14	6	20	10	23	9	23	7	10	9	11	9	12	6	10	7	8	9	18
	5	10	5	8	5	8	10	19	9	21	7	18	20	31	6	7	6	9	9	15	8	16	8	9	12	16
	7	7	5	7	7	12	6	11	5	21	7	23	12	27	4	6	5	10	5	9	6	9	9	13	10	13
	7	10	6	10	7	8	10	19	6	19	8	24	16	27	5	8	7	11	7	14	7	13	8	12	13	25
	7	13	7	9	7	11	15	29	11	24	10	22	30	45	7	11	6	9	14	27	8	15	7	7	16	41
	10	13	7	8	9	16	15	26	9	19	12	26	25	39	6	7	6	12	12	20	7	11	9	11	13	33
	13	13	6	8	10	20	13	24	8	17	19	32	19	33	5	7	11	21	11	18	6	11	12	24	12	22
4	2	2	4	5	5	6	4	7	5	14	6	14	6	15	6	5	5	7	6	6	4	5	5	5	5	7
	2	5	5	8	4	7	8	15	8	16	7	14	17	28	5	7	6	5	7	11	7	16	5	9	11	27
	2	7	4	7	4	4	12	22	9	16	5	12	29	36	6	8	6	7	10	19	8	17	4	5	16	39
	5	7	4	7	8	11	11	18	8	17	10	18	23	31	5	5	6	7	10	15	7	9	10	12	9	20
	5	10	8	11	5	9	25	41	12	17	7	17	47	17	7	11	5	7	19	29	10	17	6	7	23	49
	7	7	5	6	10	15	10	15	6	12	20	26	20	26	3	5	7	12	11	13	9	9	12	18	13	20
	7	10	6	9	8	11	27	37	8	16	15	21	40	46	7	7	9	13	18	23	9	13	10	12	25	40
	7	13	9	12	7	10	39	55	14	18	10	20	68	68	8	12	6	9	29	44	9	18	9	9	35	58
	10	13	6	6	13	20	29	42	8	14	26	31	61	62	5	6	12	17	22	36	7	11	14	19	30	47
	13	13	6	7	24	34	28	43	7	10	49	53	49	52	5	7	20	30	21	32	4	7	24	37	27	37
2	2	2	5	6	5	8	7	8	6	8	8	12	10	12	6	5	6	8	4	7	5	5	6	8	5	7
	2	5	7	9	6	6	19	30	16	17	6	11	51	54	7	11	7	8	19	27	6	12	5	6	21	41
	2	7	13	17	4	6	45	58	24	26	5	9	77	77	9	12	5	4	35	49	14	26	4	4	43	68
	5	7	7	9	9	13	34	44	11	16	24	25	64	63	6	7	10	12	27	34	6	8	10	14	34	47
	5	10	16	21	7	10	66	79	25	29	14	21	93	29	12	16	8	10	51	66	12	23	10	9	63	79
	7	7	5	5	24	33	25	35	5	7	47	51	49	50	5	6	19	25	21	28	7	7	25	35	25	33
	7	10	8	10	16	21	56	71	15	19	30	34	88	87	7	10	13	19	44	56	7	9	8	24	52	68
	10	10	4	4	39	52	45	56	5	8	76	77	73	75	4	5	36	46	33	45	7	8	40	52	43	55

properties of the individual procedures are discussed in the light of the overall results from the simulation; they are available on request. We give the interpretation on the basis of differences in CV's:

I. Both CV's are identical: P_1 achieves its nominal level rather well for all simulation parameters (see also part I). In many cases, however, P_2 and P_3 show insufficient results particularly if the size is large and the sampling distribution is skew; then they exceed their nominal level considerably. The procedures P_4, P_5, P_6 and P_7 are far away from their nominal level unless both CV's are small. To illustrate an actual level of about 50% one could say that the outcome of a method comparison can be obtained by tossing a coin.

II. Both CV's are approximately identical, say $1 < \frac{CV_y}{CV_x} < 1.5$. Here P_1 is the best of all procedures in meeting the nominal level for all simulation parameters, but it does occasionally exceed $\gamma = 5\%$. If the size c is large and the sampling distribution is skew then P_2 and P_3 exceed their nominal level considerably. If the size c is small the actual levels of P_1, P_2, P_3 are essentially higher than in case I. The actual levels of P_6 and P_7 are higher, and those of P_4 and P_5 are lower when compared with case I.

III. Both CV's are rather different, say $1.5 < \frac{CV_y}{CV_x} < 2.5$: Then P_4 is the procedure with the best result but it can also exceed γ. However, if both CV's are less than 7% and the sampling distribution is uniform then P_1 also meets the level γ rather well in contrast to P_2 and P_3. P_5 has a higher level than P_4. In comparison with case II the levels of P_4 and P_5 are decreased whereas those of P_6 and P_7 are further increased.

IV. Both CV's are essentially different, say $\frac{CV_y}{CV_x} > 2.5$: Then P_4 achieves $\gamma = 5\%$, whereas the level of P_5 is in most of the cases higher. P_1, P_2, and P_3 may also exceed $\gamma = 5\%$ considerably. The levels of P_6 and P_7 can go as far as 100%.

These results are plausible: If both CV's are very different then X might be considered to be free of random variation relative to Y, so that P_4 and P_5 will be appropriate; in contrast P_6 and P_7 are completely inappropriate. P_4 is superior to P_5 since P_4 is distribution-free. If, however, both CV's are identical then no variable can be assumed to be free of random variation, and P_1, P_2 or P_3 will be applicable. P_1 is superior to P_2 and P_3 since it is not based on any distributional assumptions.

It can be seen that the actual level of P_1 is more independent of the sampling distribution than that of P_2 or P_3, especially in the case of a large c which is typical for most of the applications. It is obvious from the results that it does not make sense to perform both P_4 and P_6 or P_5 and P_7 since it is likely that the test results will contradict each other.

4.2 The bias of the slope estimators if no extreme values are present

Again the sample size is restricted to n = 40. The resulting bias is accurate up to ± 1%. Table 2 shows the bias of $b_1, \ldots, b_7$ in per.cent of $\beta = 1$ in correspondence to table 1.

Procedures P_1, P_2 and P_3 show the same degree of bias if no extreme values are present. If any bias occurs then b_1, b_2 and b_3 overestimate β whereas b_4 and b_5 underestimate β. b_6 and b_7 underestimate β as well; however, in table 2 the bias of $\frac{1}{b_6}$ and $\frac{1}{b_7}$ is given in order to demonstrate that b_4 and b_6 or b_5 and b_7 are quite different and that they do not correspond to each other. b_4 and b_5 have a similar bias, if any, and the same holds true for b_6 and b_7. The bias increases if c becomes smaller or the CV's are increased. As before we discuss four cases:

I. $CV_x = CV_y$: The slope estimators of the invariant procedures P_1, P_2 and P_3 can be judged to be unbiased. The other procedures, however, may produce rather heavily biased results.

II. $1 < \frac{CV_y}{CV_x} < 1.5$: b_1, b_2 and b_3 often are nearly unbiased. But they show a bias if the size c is small or if the sampling distribution is skew.

III. $1.5 < \frac{CV_y}{CV_x} < 2.5$: Here b_4 and b_5 are the least biased estimators. If both CV's are less than 7% and if the sampling distribution is uniform then the invariant procedures also have only a small bias.

IV. $\frac{CV_y}{CV_x} > 2.5$: b_4 and b_5 do not show an appreciable bias whereas b_1, b_2 and b_3 may be strongly biased; b_6 and b_7 are highly biased.

If an estimator is biased the corresponding actual test level is increased. The inverse, however, is not true. The test level may be highly increased even though the estimator is unbiased – caused by a violation of the corresponding distributional assumptions. Therefore P_1 with no distributional assumptions is superior to P_2 and P_3, and P_4 is superior to P_5 even if extreme values do not occur.

Tab. 2. Bias in % of $\beta = 1$ for n = 40; no extreme values are present.

Procedures see legend to table 1.

			Samples: uniform Errors: normal							Samples: skew Errors: normal							Samples: uniform Errors: mixture of normals							Samples: uniform Errors: skew with kurtosis						
			Invariant			Not invariant				Invariant			Not invariant				Invariant			Not invariant				Invariant			Not invariant			
c	CV_x (%)	CV_y (%)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_1	P_2	P_3	P_4	P_5	P_6	P_7
8	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	5	0	0	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0	1	0	1	1	−1	−2	−1	−1
	2	7	0	1	1	0	0	1	2	1	1	1	0	0	2	3	0	1	1	0	0	1	2	0	2	2	0	1	1	3
	5	7	0	1	1	0	−1	1	2	0	0	0	−1	−2	2	2	0	0	0	−1	−1	1	1	0	1	1	0	0	1	3
	5	10	1	1	1	−1	−1	2	4	2	2	2	−1	−1	5	5	0	1	1	−1	−1	2	3	1	3	3	0	1	2	5
	7	7	0	0	0	−1	−2	1	2	1	0	0	−2	−2	3	3	0	0	0	−1	−1	1	1	0	0	0	−1	−2	1	2
	7	10	0	1	1	−1	−2	2	4	1	0	0	−2	−3	4	4	0	0	0	−1	−2	1	2	1	2	2	−1	−1	2	5
	7	13	2	3	3	−1	−2	4	6	3	4	4	−2	−2	10	10	1	2	2	−1	−1	3	5	1	4	4	−1	0	4	9
	10	13	1	1	1	−2	−4	4	6	2	2	2	−4	−5	9	9	0	1	1	−2	−3	3	5	1	2	2	−2	−3	4	8
	13	13	0	0	0	−4	−6	4	6	1	1	1	−8	−8	9	10	0	0	0	−3	−5	3	5	0	0	1	−4	−6	3	8
4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	5	0	1	1	0	−1	1	2	2	1	1	0	0	4	4	1	1	1	0	0	1	2	1	2	2	0	1	2	3
	2	7	1	2	2	0	0	3	4	3	3	3	0	−1	8	8	1	2	2	0	0	2	3	1	4	4	0	1	3	7
	5	7	0	1	1	−2	−2	3	4	2	2	2	−4	−4	8	8	0	1	1	−2	−2	2	3	0	1	1	−2	−2	3	4
	5	10	3	4	4	−1	−2	8	10	6	5	5	−4	−4	16	15	2	2	2	−1	−2	5	6	2	5	5	−2	−1	8	11
	7	7	0	0	0	−4	−4	3	4	0	0	0	−7	−7	9	8	0	0	0	−2	−3	2	3	0	1	1	−3	−3	3	5
	7	10	2	3	3	−2	−4	8	10	4	3	4	−7	−7	16	15	2	2	2	−2	−3	5	6	2	3	4	−3	−3	8	9
	7	13	4	5	6	−3	−4	12	16	9	8	9	−7	−7	28	25	3	4	4	−3	−3	10	11	4	7	7	−3	−3	12	16
	10	13	2	2	3	−6	−9	12	15	6	5	6	−13	−13	28	27	1	2	2	−5	−6	9	11	2	3	4	−6	−7	12	15
	13	13	2	3	3	−7	−9	12	15	0	0	0	−22	−20	28	25	0	0	0	−8	−11	10	12	0	0	0	−10	−13	11	15
2	2	2	0	0	0	−1	−1	1	1	0	0	0	−2	−2	2	2	0	0	0	0	−1	1	1	0	0	0	−1	−1	1	1
	2	5	2	3	3	−1	−1	5	6	6	6	6	−2	−1	15	14	2	2	2	0	−1	4	5	2	4	4	−1	0	5	8
	2	7	5	7	7	0	0	12	14	12	10	11	−2	−2	28	25	4	5	5	0	0	9	11	5	8	8	0	1	11	16
	5	7	3	4	4	−4	−6	12	14	6	5	6	−12	−10	28	25	2	2	2	−4	−5	9	10	3	4	4	−5	−6	11	13
	5	10	9	10	11	−4	−5	23	28	19	15	19	−13	−11	59	49	7	7	8	−4	−5	18	20	9	11	13	−5	−5	23	30
	7	7	0	0	0	−10	−12	11	14	0	0	0	−22	−20	28	25	1	1	1	−7	−9	9	10	0	0	0	−10	−12	11	13
	7	10	6	6	7	−10	−12	23	27	14	11	15	−21	−19	61	52	4	4	5	−8	−10	18	19	5	6	7	−11	−12	23	27
	10	10	1	1	1	−18	−21	25	28	0	0	0	−37	−34	59	52	0	0	0	−15	−17	18	20	0	1	1	−18	−20	25	28

The bias for the reciprocal values of the slope estimates is given in case of P_6 and P_7.

4.3 The influence of extreme measurement values

We now consider the case that some of the measurement values of Y differ distinctly from the corresponding values of X; the sample size is n = 40. 5% of the measurement values are systematically biased in the following manner: If the expected values y_i^* are sorted to $y_{(1)}^* \leq \cdots \leq y_{(40)}^*$, then $y_{(30)}^*$ is decreased either by 50% or by 50% of the amount $1 - \frac{1}{c}$ whichever is the smaller. In the same way $y_{(40)}^*$ is increased by 50% or by 50% of the amount $1 - \frac{1}{c}$. Clearly, $y_{(40)}^*$ lies outside the common range of both methods; but since the actual range of Y must be larger when extreme values are present, it is a realistic model. Table 3 shows the probability of a false positive test result, i.e. the actual test level, and table 4 gives the bias of $b_1, \ldots, b_7$ in per cent of $\beta = 1$.

The actual level achieved by P_1 is equal to or slightly higher than $\gamma = 5\%$, if both CV's are approximately identical – provided the sampling distribution is uniform. If the sampling distribution is skew and the size c not large then the actual level of P_1 may be considerably higher than 5%. Procedure P_4 meets its nominal level very well provided that both CV's are rather different. b_1 tends to be biased particularly if the sampling distribution is skew and c is not large.

The actual level of P_2 and P_3 exceeds $\gamma = 5\%$ by far, and is substantially higher than that of P_1 for all simulation parameters. If the sampling distribution is skew and the precision of both methods is high it can go up to 100%: here P_2 and P_3 would declare both methods to be significantly different even though $\beta = 1$ is true. Both b_2 and b_3 are biased particularly if the sampling distribution is skew.

Now we consider the influence of the number of extreme values and their location (below or above the

Tab. 3. Probability of a false positive test result for n = 40 (γ = 5%); two extreme values are present.

Procedures see legend to table 1.

			Samples: uniform Errors: normal						Samples: skew Errors: normal					
			Invariant		Not invariant				Invariant		Not invariant			
c	CV_x (%)	CV_y (%)	P_1	P_2 P_3	P_4	P_5	P_6	P_7	P_1	P_2 P_3	P_4	P_5	P_6	P_7
∞	2	2	4	22	3	0	8	100	7	100	6	91	11	100
	2	5	7	38	6	0	14	99	6	100	5	74	19	100
	2	7	9	42	5	0	19	99	8	100	4	63	27	100
	5	7	6	35	5	0	15	94	6	99	3	46	22	100
	5	10	9	41	4	1	24	93	14	96	5	45	40	100
	7	7	7	26	3	1	14	90	5	97	4	36	21	100
	7	10	7	32	3	1	19	89	8	91	6	31	33	99
	7	13	8	34	5	1	25	87	11	87	4	28	49	99
	10	13	7	28	5	1	25	83	8	81	8	16	44	98
	13	13	9	24	8	2	25	82	6	69	8	9	40	96
4	2	2	7	15	3	0	13	100	8	100	3	13	20	100
	2	5	8	28	4	0	20	98	14	100	4	22	44	100
	2	7	8	36	4	0	27	97	23	98	5	23	65	100
	5	7	7	24	3	0	27	90	12	90	4	10	56	100
	5	10	14	34	5	1	42	94	26	88	3	16	82	100
	7	7	6	20	5	0	23	86	8	79	9	6	50	99
	7	10	8	27	4	0	40	89	18	77	6	6	71	100
	7	13	12	33	5	1	56	94	25	81	6	7	90	100
	10	13	8	22	9	4	49	87	16	64	15	4	83	97
	13	13	5	15	12	11	45	83	9	44	33	11	77	95
2	2	2	5	17	2	0	15	99	8	100	4	7	29	100
	2	5	10	37	2	0	40	96	31	97	3	11	81	100
	2	7	20	47	4	2	66	96	48	94	3	11	95	100
	5	7	11	26	5	2	54	91	23	75	10	3	90	100
	5	10	18	39	5	4	79	98	49	83	8	4	99	100
	7	7	4	12	16	10	43	81	7	43	36	9	80	97
	7	10	11	25	10	7	73	94	26	62	23	7	97	100
	10	10	5	10	30	32	61	85	6	27	60	37	91	97

Tab. 4. Bias in % of β for n = 40; two extreme values are present.
Procedures see legend to table 1.

			Samples: uniform Errors: normal							Samples: skew Errors: normal						
			Invariant			Not invariant				Invariant			Not invariant			
c	CV_x (%)	CV_y (%)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_1	P_2	P_3	P_4	P_5	P_6	P_7
∞	2	2	0	10	11	0	4	0	16	1	24	25	0	16	1	32
	2	5	1	11	11	0	5	1	16	2	25	26	1	16	3	33
	2	7	1	11	12	1	5	2	18	2	25	27	1	16	5	33
	5	7	2	11	11	0	4	3	18	2	23	26	0	15	5	33
	5	10	2	12	12	1	4	5	19	4	25	28	0	16	10	37
	7	7	1	10	10	0	3	3	18	2	23	25	− 1	14	7	33
	7	10	2	11	11	0	3	5	19	3	24	27	− 2	13	10	35
	7	13	3	12	13	− 1	2	7	22	6	26	28	− 1	12	13	39
	10	13	3	11	12	− 1	1	8	22	5	25	28	− 3	11	15	41
	13	13	3	10	11	− 3	− 1	8	23	4	23	27	− 7	7	16	41
4	2	2	1	11	12	0	4	1	19	1	26	29	0	15	3	39
	2	5	2	11	12	0	3	3	20	4	27	30	1	15	9	41
	2	7	3	13	14	1	3	7	22	8	28	32	1	15	15	45
	5	7	3	11	13	− 1	1	7	22	7	27	31	− 2	11	16	45
	5	10	5	12	14	0	1	11	25	13	30	36	− 1	12	28	52
	7	7	2	10	11	− 2	− 1	8	22	5	24	28	− 6	7	18	43
	7	10	5	12	13	− 2	− 1	12	25	10	27	32	− 6	8	28	49
	7	13	7	14	17	− 2	− 1	18	33	16	31	39	− 6	7	41	59
	10	13	6	12	14	− 6	− 6	19	33	13	27	35	−13	1	45	59
	13	13	4	10	12	− 9	− 9	19	33	8	21	28	−21	− 8	47	59
2	2	2	1	10	11	0	2	3	19	3	27	31	− 2	13	7	43
	2	5	4	13	14	0	3	10	23	13	31	36	0	12	25	52
	2	7	9	16	18	1	2	16	32	22	35	43	0	12	41	61
	5	7	7	13	16	− 4	− 3	18	32	16	29	37	−11	2	45	61
	5	10	12	17	22	− 5	− 4	30	43	32	37	52	− 9	2	79	85
	7	7	4	10	11	− 9	− 9	18	32	10	22	29	−21	− 8	49	61
	7	10	9	15	19	− 9	− 9	32	45	24	30	44	−20	− 7	79	82
	10	10	5	9	12	−18	−18	32	45	10	19	31	−36	−23	82	82

regression line). Table 5 shows the simulated cases, where the extreme values are obtained as described above. Table 6 states the respective probability of a false positive test result and the bias of procedures P_1, P_2 and P_3 if both CV's are 7% (in this constellation procedures P_4, P_5, P_6, P_7 are completely inadequate).

Tab. 5. Distribution of the extreme values for n = 40.

Number of extreme values	% of n	Position in the sort sequence of the y_i	Location with respect to regression line
1	2.5	40	above
2	5.0	30	below
		40	above
3	7.5	30	below
		35	above
		40	above

Tab. 6. Actual probability of a false positive test result (γ = 5%) and bias in % of β = 1 for n = 40 and $CV_x = CV_y$ = 7%, depending on the number and distribution of extreme values.
Procedures see legend to table 1.

	Extreme values		Samples: uniform Errors: normal						Samples: skew Errors: normal					
			Actual level			Bias			Actual level			Bias		
c	Number	% of n	P_1	P_2 P_3		P_1	P_2	P_3	P_1	P_2 P_3		P_1	P_2	P_3
∞	0	0	5	7		0	0	0	5	21		1	0	0
	1	2.5	9	57		2	11	11	5	97		2	22	24
	2	5	7	26		1	10	10	5	97		2	23	25
	3	7.5	11	77		3	16	18	12	99		4	29	33
4	0	0	5	6		0	0	0	6	12		0	0	0
	1	2.5	8	33		2	10	10	7	81		3	22	25
	2	5	6	20		2	10	11	8	79		5	24	28
	3	7.5	10	62		5	18	21	15	97		8	35	41
2	0	0	5	5		0	0	0	5	7		0	0	0
	1	2.5	6	16		3	9	10	7	40		4	18	24
	2	5	4	12		4	10	11	7	43		10	22	29
	3	7.5	10	34		8	16	20	17	74		16	32	45

The probability of a false positive test result obtained by P_1 can be higher than $\gamma = 5\%$ and b_1 may be biased. P_2 and P_3, however, exceed the nominal level to a much greater extent and show a considerably greater bias. The extreme point located at the upper boundary of the range of X and above the regression line causes most of these effects whereas the extreme value located inside the range and below the regression line does not show any extra effect. An additional extreme value near the upper boundary of the range of X and above the line increases the effect.

Therefore, if there is at least one extreme value present P_1 is superior to P_2 and P_3 since the results given by P_1 are only slightly impaired. This property holds for both a uniform and a skew sampling distribution; it is a generalization of our example in part I of this paper.

4.4 Very small size of measurement range

All previous statements hold also for a very small size c, but all effects are amplified even when the precision of both methods is high. Table 7 shows the probability of a false positive test result and the bias of b_1, b_2 and b_3 if two extreme values are present (see 4.3). The interpretation is obvious.

Tab. 7. Actual probability of a false positive test result ($\gamma = 5\%$) and bias in % of $\beta = 1$ for $c = 1.25$ and $n = 40$; two extreme values are present.

Procedures see legend to table 1.

		Samples: uniform Errors: normal					Samples: skew Errors: normal				
		Actual level		Bias			Actual level		Bias		
CV_x (%)	CV_y (%)	P_1	P_2 P_3	P_1	P_2	P_3	P_1	P_2 P_3	P_1	P_2	P_3
1	1	5	20	2	10	11	9	93	5	26	31
1	2	14	37	6	13	15	30	90	17	31	38
2	2	5	14	3	9	11	9	50	8	22	30

4.5 The influence of a non-constant CV on the behaviour of the procedures

We investigated three different situations in which the variation of the CV was the same for both methods:

A constant standard deviation in the lower part (first 20%) and a constant CV in the rest of the measurement range do not influence the results.

A constant CV for 80% of the measurement range, followed by an increasing CV up to double its size, affects the results of procedure P_1 only slightly whereas the results from procedures P_2 and P_3 become seriously impaired.

A non-constant CV at both ends of the measurement range, however, does not lead to a further deterioration of the results.

4.6 The influence of the sample size on the above results

The results of our simulation study show that the estimates of the slope β do not change if the sample size of $n = 40$ is enlarged to 60 or 80; i.e. the bias – if any – is independent of n. Therefore the results obtained in 4.2 to 4.5 do not depend on the sample size.

The probability of a false positive test result is also independent of n if the corresponding estimator is unbiased. Otherwise the actual level increases if n is increased. The judgement and ranking of procedures $P_1, \ldots, P_7$, however, remain valid.

However, it must be stressed that these statements of independence of n are only valid for the average of the estimation; its precision decreases with the sample size.

4.7 Recommendations

There is no procedure which can be applied without restrictions for the statistical evaluation of a method comparison unless the appropriate experimental design is used. In particular, it has been demonstrated that the quality of statistical results from P_2 and P_3 may be impaired if the sampling distribution is not uniform. Reliable statistical results, however, can be expected from procedure P_1 even if some extreme values are present, provided that the following recommendations are observed.

- The sampling distribution should be uniform. Under the usual experimental conditions, however, it is more likely to be skew. In this case the results from P_1 are more reliable than those from the other procedures.
- If the precisions of both methods are identical or nearly identical then P_1 is the most reliable of all the procedures, even if the CV's are not constant over the measurement range.
- If the precisions of both methods are different then P_1 gives reliable results, provided that both CV's are less than 7%.

– For any other constellation of the CV's there are two possibilities:

  P_4 is the most appropriate of all, but the results are not invariant with respect to the assignment to X and Y.

  If it is required that X and Y can be interchanged, then without impairing the results an alternative experimental design is recommended: Single determinations should be performed with the more precise method (X) whereas k-fold determinations should be performed with the less precise one (Y). k is determined from

$$k \geq \left(\frac{CV_y}{f \cdot CV_x}\right)^2 \qquad \text{(Equ. 1)}$$

  where f is obtained from table 8.

Tab. 8. Factor f determining the number k of replicate measurements for the less precise method.

c	f
∞	4
4	2
2	1.5

If we take the mean of these k-fold determinations for each sample we ensure that its coefficient of variation differs from CV_x by the factor f only. Then taking as input the single values of the one method and the means of the k-fold determinations of the other method, P_1 can be used for evaluation.

5. The Sample Size

5.1 The probability of a significant test result

The power of procedure P_1 if $\beta > 1$ and $CV_x = CV_y$ is given in tables 9, 10 and 11 for different sizes c. The accuracy of the values is in the range of ± 2.

To demonstrate the dependence of the power on the simulation parameters, consider the following example: Let c be large, both CV's 7%, n = 60 and the sampling distribution skew, then the chance of a significant test result is about 40% if the true but unknown value of β is 1.06 (see tab. 9).

Tab. 9. Probability of a significant test result obtained by P_1 if the size is *large* (c = ∞) and both CV's are equal (γ = 5%).

CV (%)	n	Samples: uniform Errors: normal β							Samples: skew Errors: normal β						
		1.02	1.04	1.06	1.08	1.10	1.15	1.20	1.02	1.04	1.06	1.08	1.10	1.15	1.20
2	30	49	97	100	100	100	100	100	31	78	97	100	100	100	100
	40	65	99	100	100	100	100	100	37	88	100	100	100	100	100
	50	75	100	100	100	100	100	100	46	95	100	100	100	100	100
	60	80	100	100	100	100	100	100	53	98	100	100	100	100	100
	70	88	100	100	100	100	100	100	60	99	100	100	100	100	100
	80	92	100	100	100	100	100	100	66	100	100	100	100	100	100
	90	95	100	100	100	100	100	100	72	100	100	100	100	100	100
7	30	11	20	41	57	79	97	100	9	13	22	35	48	80	93
	40	12	27	50	70	87	100	100	9	15	28	47	60	90	99
	50	13	34	59	83	94	100	100	10	19	36	53	71	95	100
	60	14	40	67	90	97	100	100	11	22	40	62	81	99	100
	70	15	45	75	93	98	100	100	11	25	45	70	87	100	100
	80	17	49	80	97	100	100	100	12	27	52	75	90	100	100
	90	20	53	86	98	100	100	100	12	29	57	80	94	100	100
13	30	6	10	15	23	30	58	79	6	7	10	16	21	37	58
	40	7	11	19	34	42	70	91	7	10	12	18	25	49	66
	50	7	13	24	38	51	83	96	7	10	17	22	31	58	78
	60	8	15	29	44	59	90	98	8	11	19	26	36	67	85
	70	8	18	32	48	67	94	99	8	12	21	30	41	74	91
	80	8	20	34	51	76	98	100	8	13	22	33	46	79	94
	90	9	21	38	55	83	99	100	9	16	23	36	51	84	96

Tab. 10. Probability of a significant test result obtained by P_1 if the size c is *medium* (c = 4) and both CV's are equal (γ = 5%).

CV (%)	n	Samples: uniform Errors: normal β							Samples: skew Errors: normal β						
		1.02	1.04	1.06	1.08	1.10	1.15	1.20	1.02	1.04	1.06	1.08	1.10	1.15	1.20
2	30	20	66	92	100	100	100	100	11	35	70	84	95	100	100
	40	28	76	98	100	100	100	100	16	45	77	95	100	100	100
	50	35	85	100	100	100	100	100	20	55	87	98	100	100	100
	60	41	95	100	100	100	100	100	24	61	93	100	100	100	100
	70	46	97	100	100	100	100	100	28	68	97	100	100	100	100
	80	50	98	100	100	100	100	100	30	75	100	100	100	100	100
	90	55	100	100	100	100	100	100	33	80	100	100	100	100	100
7	30	7	9	15	26	36	67	90	5	9	13	18	24	43	64
	40	8	12	22	33	50	79	97	6	9	14	21	28	50	76
	50	8	14	26	42	60	90	99	7	10	16	25	33	58	83
	60	8	17	32	50	69	96	100	7	11	18	29	40	67	90
	70	9	20	36	57	76	97	100	8	12	21	32	46	74	94
	80	9	22	41	62	82	100	100	8	13	24	36	52	81	96
	90	9	24	45	66	83	100	100	8	16	26	39	56	86	99
13	30	5	6	9	14	13	29	46	5	7	7	8	13	20	27
	40	5	7	10	17	20	38	57	5	8	9	10	16	27	38
	50	5	7	12	19	25	47	68	6	8	11	12	18	32	46
	60	5	8	14	21	29	55	77	6	9	12	15	20	35	52
	70	5	9	16	23	34	62	83	7	9	13	17	24	38	59
	80	6	10	18	26	38	69	87	7	10	14	19	26	43	65
	90	6	11	20	29	42	72	91	7	10	16	21	29	48	69

Tab. 11. Probability of a significant test result obtained by P_1 if the size c is *small* (c = 2) and both CV's are equal (γ = 5%).

CV (%)	n	Samples: uniform Errors: normal β							Samples: skew Errors: normal β						
		1.02	1.04	1.06	1.08	1.10	1.15	1.20	1.02	1.04	1.06	1.08	1.10	1.15	1.20
2	30	10	26	50	77	92	100	100	7	16	24	43	58	86	97
	40	12	37	66	86	97	100	100	8	19	34	55	73	94	99
	50	14	42	73	91	99	100	100	9	22	41	66	82	98	100
	60	16	49	81	98	100	100	100	11	24	49	72	87	100	100
	70	18	56	86	99	100	100	100	12	27	57	77	91	100	100
	80	21	63	91	100	100	100	100	13	33	62	84	96	100	100
	90	23	69	93	100	100	100	100	14	35	65	88	97	100	100
7	30	5	7	9	13	16	37	49	5	6	7	9	10	16	25
	40	5	7	11	17	21	41	60	5	7	8	11	14	20	38
	50	5	8	12	20	24	48	72	5	8	10	13	18	26	44
	60	6	9	14	22	31	55	79	6	8	11	16	21	32	53
	70	6	9	16	24	36	65	85	6	9	12	17	23	37	59
	80	7	10	17	27	41	72	92	6	9	13	18	27	42	67
	90	7	11	18	29	45	78	94	6	10	14	19	29	49	73
10	30	5	5	6	9	13	20	30	5	5	6	7	9	15	21
	40	5	6	7	11	15	25	38	5	5	6	8	11	18	27
	50	5	8	8	12	18	31	47	5	6	7	9	12	22	32
	60	6	8	9	14	20	37	53	5	6	7	11	14	25	39
	70	6	9	11	16	22	44	62	6	6	8	13	16	29	43
	80	6	9	12	18	26	49	70	6	7	9	14	18	33	47
	90	6	9	14	20	29	52	73	6	7	9	15	20	37	51

From the tables the following properties can be deduced:

- The power decreases with the size c. For the above example the power goes down to 11% if c = 2 (see tab. 11). This is intuitively clear since the estimation is more precise if the size c is large and a precise estimator b_1 of β causes a high power.
- The power decreases with increasing CV's. If both CV's are 2% then the power is as high as 100% whereas if both CV's are 13% the power goes down to 19% (see tab. 9).
- The power increases with the value of β. It reaches about 100% if β = 1.15. Obviously the more both methods differ from each other the better the difference can be detected (see tab. 9).
- The power increases with increasing n. If n = 90 then the power goes up to 57%, but the power is 22% only if n = 30 (see tab. 9). The power approaches 100% if n is large enough.
- The uniform sampling distribution leads to a higher power than the skew sampling distribution. In the above example, the power is 67% in the case of a uniform sampling distribution (see tab. 9).

Analogous statements hold if the CV's are not equal. They also remain valid if the CV's are very different and the evaluation is done by procedure P_4 (see 4.7).

Therefore, the reliability of the result when testing the hypothesis β = 1 depends not only on n but also on the other conditions, i.e. on size c, CV_x and CV_y, true value of β, and the sampling distribution. A fixed sample size independently of these conditions cannot be recommended.

The consequences of an inappropriate sample size can be demonstrated by the above example. Let c = ∞, $CV_x = CV_y = 7\%$, and let the sampling distribution be uniform. Assume that a deviation of 8% or more from β = 1 is relevant, i.e. $\beta_{rel} = 1.08$. If n = 30, then the power is 57% if the true but unknown β is 1.08; however, the power is 79% if β = 1.10 holds (see tab. 9). If therefore the unknown β were 1.08 or a little bit larger, this deviation would not be detected with a sufficient probability (≥80%), though the value of β_{rel} would be exceeded. On the other side let n = 90. Then the power is 86% even if β = 1.06 holds; i.e. this deviation would be detected as significant with a high probability even though it would not be relevant. A reasonable sample size, however, would be n = 50. If the unknown β exceeds the threshold β_{rel} then the probability of a significant test result is at least 83%; but for $\beta < \beta_{rel}$ this probability is lower. The sample sizes given in the next chapter are found in this way (under the assumption of identical CV's).

If the methods have different CV's then the power of the test will be different depending on whether β or 1/β is greater than 1. As a consequence different sample sizes are also required. Since it is usually not known before the evaluation whether the outcome is β > 1 or 1/β > 1 we recommend the use of the larger sample size; the tabulated sample sizes take this into account.

5.2 Determination of the sample size

The common size c of the measurement range of both methods is known before the beginning of an experiment. The same is true with the precisions of both methods. If the coefficients of variation do not vary too much over their concentration range the precisions may be expressed by CV_x and CV_y. The investigator has to define by β_{rel} a relevant difference to β = 1.

If both precisions do not differ by more than the factor f, evaluation should be carried out by procedure P_1; table 12 gives the respective sample sizes if the sampling distribution is uniform; table 13 gives the appropriate sample sizes in the case of a skew sampling distribution. If n is determined in this way then the probability of detecting a relevant deviation from β = 1 by procedure P_1 is about 80% or more.

It is evident that a skew sampling distribution requires a larger sample size than a uniform sampling distribution; the skew distribution as described, requires approximately twice the sample size of a uniform distribution. On the other hand, to obtain a uniformly distributed sample one needs to carry out more than the required number of determinations to achieve such a distribution. To aim for the sample size of a skew distribution may be the easier solution; the properties of procedure P_1 are still acceptable in this case.

If the precisions of both methods are different but both are less than 7% then the evaluation can be carried out by P_1 and n is obtained from table 12 or 13.

If the precisions of both methods are different and at least one CV is larger than 7% then the evaluation can be performed by P_1 provided that k-fold determinations have been carried out for each sample of the less precise method. The value of k is obtained from equation (1) and n from table 12 or 13.

The assumptions with regard to the shape of the skew sampling distribution and to normally distrib-

Tab. 12. Proposed sample sizes for procedure P_1 if the sampling distribution is *uniform* (γ = 5%; power at least 80%).

			β_{rel} / $1/\beta_{rel}$							
c	CV_x (%)	CV_y (%)	1.02 0.98	1.04 0.96	1.06 0.94	1.08 0.93	1.10 0.91	1.12 0.89	1.15 0.87	1.20 0.83
∞	2	2	60	–	–	–	–	–	–	–
	2	5	+	60	30	–	–	–	–	–
	5	2	+	60	30	–	–	–	–	–
	5	5	+	90	40	30	–	–	–	–
	5	7	+	+	70	40	30	–	–	–
	7	5	+	+	70	40	30	–	–	–
	7	7	+	+	80	50	30	–	–	–
	7	10	+	+	+	90	55	40	30	–
	10	7	+	+	+	90	55	40	30	–
	10	10	+	+	+	90	60	45	30	–
	10	13	+	+	+	+	90	70	45	–
	13	10	+	+	+	+	90	70	45	–
	13	13	+	+	+	+	90	75	50	30
4	2	2	+	45	–	–	–	–	–	–
	2	5	+	+	80	45	30	–	–	–
	5	2	+	+	80	45	30	–	–	–
	5	5	+	+	+	65	45	–	–	–
	5	7	+	+	+	+	70	45	35	–
	7	5	+	+	+	+	70	45	35	–
	7	7	+	+	+	+	75	55	40	–
	7	10	+	+	+	+	+	+	70	40
	10	7	+	+	+	+	+	+	70	40
	10	10	+	+	+	+	+	+	70	40
	10	13	+	+	+	+	+	+	+	70
	13	10	+	+	+	+	+	+	+	70
	13	13	+	+	+	+	+	+	+	70
2	2	2	+	+	60	35	–	–	–	–
	2	5	+	+	+	+	+	75	40	–
	5	2	+	+	+	+	+	75	40	–
	5	5	+	+	+	+	+	90	60	40
	5	7	+	+	+	+	+	+	90	60
	7	5	+	+	+	+	+	+	90	60
	7	7	+	+	+	+	+	+	+	65
	7	10	+	+	+	+	+	+	+	+
	10	7	+	+	+	+	+	+	+	+
	10	10	+	+	+	+	+	+	+	+

– represents sample sizes <30; the value for β_{rel} is rather large in respect to the CV's.
+ represents sample sizes >90.

uted measurement errors have a serious influence on the appropriate sample size; n would have to be increased if the sampling distribution becomes more skew. The distribution of measurement errors is generally not known; therefore we refer to a "standard" distribution. It must be noted here that the distribution of measurement errors may affect the power, but the chance of a false positive test result (i.e. if $\beta = 1$ is true) remains about 5%.

6. Conclusion

Statistical evaluation of a method comparison by linear regression is completely inappropriate since it is rather likely that it produces misleading results. The results of the principal component analysis (procedure P_3) are more reliable, and those of the standardized principal component analysis (procedure P_2) even better. But both principal component procedures show important drawbacks if the sampling distribution is skew, if the CV's are not constant over the measurement range, or if extreme values are present. Because of its robustness the new procedure P_1 can also cope with those situations; it is the regression procedure we recommend for the statistical evaluation of method comparison studies.

Acknowledgement

The authors are very grateful to *W. Bardorff, R. Bender, H.-G. Eisenwiener, R. Spaethe, W. Specht* and *E. Völkert* for their valuable comments and Prof. *B. Schneider* for his helpful suggestions.

Tab. 13. Proposed sample sizes for procedure P_1 if the sampling distribution is *skew* (γ = 5%; power at least 80%).

			β_{rel}							
			$1/\beta_{rel}$							
c	CV_x (%)	CV_y (%)	1.02 0.98	1.04 0.96	1.06 0.94	1.08 0.93	1.10 0.91	1.12 0.89	1.15 0.87	1.20 0.83
∞	2	2	+	30	–	–	–	–	–	–
	5	5	+	+	80	45	35	–	–	–
	7	7	+	+	+	90	60	45	30	–
	10	10	+	+	+	+	+	80	55	35
	13	13	+	+	+	+	+	+	80	50
4	2	2	+	90	40	–	–	–	–	–
	5	5	+	+	+	+	85	65	40	–
	7	7	+	+	+	+	+	+	80	45
	10	10	+	+	+	+	+	+	+	80
	13	13	+	+	+	+	+	+	+	+
2	2	2	+	+	+	75	50	35	–	–
	5	5	+	+	+	+	+	+	+	80
	7	7	+	+	+	+	+	+	+	+
	10	10	+	+	+	+	+	+	+	+

– represents sample sizes <30.
+ represents sample sizes >90.

References

1. Passing, H. & Bablok, W. (1983) J. Clin. Chem. Clin. Biochem. *21*, 709–720.
2. Feldmann, U., Schneider, B., Klinkers, H. & Haeckel, R. (1981) J. Clin. Chem. Clin. Biochem. *19*, 121–137.
3. Averdunk, R. & Borner, K. (1970) J. Clin. Chem. Clin. Biochem. *8*, 263–268.
4. Theil, H. (1950) Proc. Kon. Akad. v. Wetensch. AS 3 Part I 386–392.
5. Ostle, B. (1963) Statistics in Research, The Iowa State University Press, Ames
6. Bartley, P., Fox B. L. & Schrage, L. E. (1983) A Guide to Simulation, Springer Verlag, New York, Heidelberg.
7. Eisenwiener, H.-G., Bablok, W., Bardorff, W., Bender, R., Markowetz, D., Passing, H., Spaethe, R. & Specht, W. (1983) Lab. Med. *7*, 273–281.
8. Passing, H. (1981) J. Clin. Chem. Clin. Biochem. *19*, 1145–1154.
9. Michotte, Y. (1978) Evaluation of precision and accuracy – comparison of two procedures, In: Evaluation and optimization of laboratory methods and analytical procedures (Massart, D. L., Dijkstra, A. & Kaufmann, L., eds.) Elsevier, Amsterdam, Oxford, New York.

Dr. H. Passing
Abtlg. für Praktische Mathematik
Hoechst AG
D-6230 Frankfurt 80

W. Bablok
Allgemeine Biometrie
Boehringer Mannheim GmbH
D-6800 Mannheim 31